Reclassifying the Classified: Practical Considerations in Mining HR Data

Mehmet Gençer , Savaş Yıldırım
Istanbul Bilgi University

Introduction

This research study is driven by the following problems presented to us by a human resources consulting company:

1. The costs related to personnel turnover, ie. existing personnel leaving the organization.
2. The costs related to less-than-expected performance of new hires.
3. Financial and motivational costs due to promotion of inappropriate candidates to management ranks.
4. Financial and motivational costs due to repositioning of internal talent into open positions.
5. Motivational costs on others due to promotion of personnel who is ill-qualified for the position.

This set of problems represent tangible issues which the human resources experts encounter on a daily basis. An essential concern is to avoid errors due to subjectivity and similar shortcomings. The desired solution involves presenting the decision makers with some automated assessments. There are certain commonalities, some more obvious than others, among these problems stated.

Here we discuss these problems at the intersection of two perspectives:

1. *The domain specific knowledge about the the general and specific human resources problems addressed in this research.* This perspective not only allows to capture the commonalities, but also helps us to lay out the structure of the problem.
2. *Generic statistical learning and modeling methodologies.* Each of these statistical apparatus may have both strengths and weaknesses in the face of problem structure in hand.

Domain specific aspects

The human resources (HR) function aims to first determine the numbers and skill-sets of people needed by a business organization, then to find and recruit such people, and finally assess and train those who are already part of the organization. As a separate issue determining who to lay off, when necessary, is also the responsibility of HR function.

Since human and organizational resources are intangible and hard-to-imitate competencies of firms (Barney and Hesterley, 2009), it is a very critical task to maintain and improve the organizational atmosphere for HR function. On the one hand, HR undertakes the responsibility of finding and recruiting ‘compatible’ people who can adapt to organizational culture quickly and easily. On the other hand it is equally important to bring in both skills and change initiatives which the current organization lacks. These concerns apply equally to several HR duties such as recruitment, promotion, repositioning, and so on. Thus the HR function must consider both *supplementary* and *complementary* fit when considering addition or removal of personnel to a team or department (Malinowski et al., 2008).

The issue in hand is that such assessments and decisions are not only costly, but also subjective. It only gets worse in the face of accumulation of relevant data collected within organizations. In the domain of HR, use of computational approached needs to go beyond exploratory statistics which shows only *what is* (Long and Troutt, 2003), and find more about how and why. In such attempt, however, it is often necessary to combine both domain specific knowledge and data mining techniques. The need is most obvious in cases where statistical correlation is not sufficient to make the right decision regarding the causality relation (Feelders et al.,2000), and a wrong decision would be extremely costly to the business.

For this reason let us attempt a closer consideration of the structure of the particular problems in hand. One usually finds a plethora of HR data available at organizations. These data includes variables spanning both (i) employee data (e.g. age, gender, marital status, experience and skill-set, etc.) and (ii) job data (epmloyee’s job title, salary, training records, promotional history, performance appraisal ratings, etc.).

The particular and complete set of the variables, as well as their timespan and granularity, may change from one organizational setting to the other. Despite such variations, however, we can make the following generalizations and simplifications about the problems:

- The specific problems listed at the beginning all have a common core task: sorting/ranking candidates with respect to how well their profile matches a job specification in hand.
- The job specifications may be inherently less reliable when it comes to higher ranks (ie. managerial positions), which we need to further check with field experts. Nevertheless, any application of data mining methods to HR must keep in mind that the reliability of methods can only get as relible as the specifications available. Similar is the case for more intricate skills whose reliable assessment is costly. Overall, it is the task of data scientist to provide a proper assessment of reliability of the prescribed methods, conditional to such situational parameters.
- One may safely assume that the supplementarity/complementarity duality discussed by Malinowski et al (2008) can be reduced to a single criteria of `fitness`. Indeed, the solution proposed in their research is based on recommender system methods essentially work against a single criteria. Therefore, regardless of the particular definition of fitness, we assume that its definition is unique and not dualistic. However, we do consent that such a definition can be contextual. For example, the definition of fitness for

a particular position may subsume the contextual parameters regarding complementarity.

- The problem 2, which concerns external hires has the additional pitfall that the set of variables about the candidate and the set of variables about the job specification may come from different sources and thus may be incompatible. We will not discuss this issue in depth. A pragmatic, first-cut solution to the problem is confining the methods to conjunction of the two variable sets.
- In problems 1 and 2, the specification of who is likely to stay or leave the company is not given itself, but required additional data mining to be clarified.

Methodological aspects

A great stock of methodological apparatus is available today, for both classification and fitness assessment (Hastie et al, 2009;Agresti 1996). Let us first summarize the promise of the two family of methods in HR problems, then attempt an assessment of their practical strengths and weaknesses when used in HR problems in hand.

Summary of the methods

The *classification* methods provide only limited sorting/ranking of candidates since the output is merely about which class an input belongs. For example one can classify a candidate as either fit or unfit for the job specification (a binary classification). While this may seem somewhat limiting, these methods can be easier to adapt. In the HR domain, for example, one only has the logistic (true/false) information about which employee profiles have stayed with the company and which ones have left (apart from the duration of stay). In such a case classification becomes a supervised learning problem. The recommender system in Malinowski et al (ibid) for example uses a variation of classification methods. Despite its limitations, this family of methods can be a natural choice for shortlisting candidates, perhaps after some parameter tuning.

The other group, regression methods, can be used to put a scalar value on the fitness of a candidate for the specified job, which in turns results in full-scale ranking/sorting of the candidate pool. Despite its promise, or rather precisely because of it, however, this family of methods are troublesome when one has logistic information on the input side, i.e. as part of candidate profiles (Agresti, 1996; Hastie et al., 2009). In general, when using categorical inputs, it is hard to control model variance when using these methods, hence their reliability is both problematic and hard to assess.

An additional consideration for problems 1 and 2 is the need for discovering the job specifications, e.g. who is more likely to stay or leave, what is the profile of new hires who is likely to perform better, etc. These specifications needs to be found from existing data that represents past experience, by means of methods such as clustering.

One must be aware of a more general aspect of statistical methods when choosing one from either family: some methods which are non-parametric simply do not give us any information about the 'why's and 'how's of the phenomena under study. In general parametric methods (e.g. linear regression) makes assumptions about the model/structure of the phenomenon in hand. Hence when the method is worked out against data, whatever precision is there is about the model. On the other side the non-parametric methods (e.g. k-nearest neighbours) do not make such assumptions. Hence even if the method works (ie. when you work it out against data it results in good precision), it does the job (e.g. classification of which candidates are suitable for a given job spec) but it does not state anything about the reason in terms of the problem structure.

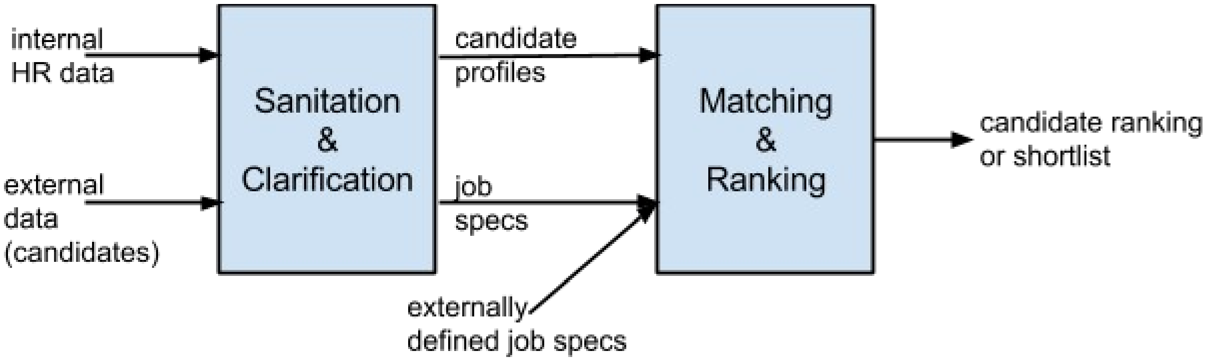
Practical considerations

Many real life problems, stuck on some inherent limitations, some of which span both family of methods. Let us summarize the potential strengths and weaknesses of these methods, relevant for our own problem set:

- The number of variables (in both input and output data sets) is simply too large for most statistical methods (Hastie et al., 2009). When the number of dimensions, n (i.e. the number of variables), is high, the samples are simply sparse in the n-dimensional space even if the number of samples, N, is high. Essentially the sampling density is $N^{1/n}$. Also most sample points are closer to some edge of the hyperspace in such a case, forcing regression methods to extrapolate, rather than intrapolate between the sample points. Hence the resulting model will have bias towards such edges.
  A common remedy to such problems is to reduce the number of dimensions, either by selecting only few variables, or applying a dimension reduction technique (e.g. those rely on latent variables, such as 'principal component' based methods.
- One needs a careful examination of outliers (in each variable in the data set) before proceeding to statistical procedures, since some procedures are more sensitive than others to outliers.
- Stability of any statistical or machine learning method depends on the reliability of the input data. In a high dimensional data set one needs further scrutiny to ensure quality of data fed to such procedures.
- Most parametric methods are based on linear models. To capture more complex relationships one needs transformation 'tricks' such as basis expansion.

A two-stage approach to solution

In approaching the solution, informed by the above analysis, we consider a two stage solution architecture as shown in the figure.


internal HR data
external data (candidates)
Sanitation & Clarification
candidate profiles
job specs
externally defined job specs
Matching & Ranking
candidate ranking or shortlist

Function of the sanitation and clarification stage is to generate job specifications and candidate profile data (whether external or internal candidates) that are clarified and compatible, and is sanitized of errors and -if necessary- outliers. This stage may necessiate using machine learning or statistical methods to determine job specifications. For example if one needs to know the specs for high performance employees in a certain role, one needs to consider applying a method such as clustering to available HR data.

Once these are ready one can applying start matching and ranking procedures. In doing so job specifications which are provided externally can enter the process, as in the case of a brand new organizational role (e.g. a company hiring a biostatistician for the first time). Depending on the problem one may use several methods ranging from non-parametric classification methods to parametric probabilistic models. Unless the target is a new organizational role, supervised learning methods is available as an option here. Choice of classification methods would result in a shortlist as the output of this stage, whereas other methods can lead to a full ranking/ordering of candidates. A combination of the two is also possible, and statistical advice is to consider the error rates or all by looking at the distributional characteristics of the particular data set in hand.

The original problems in than can now be mapped to this two stage framework roughly as shown in the table:

	inputs	stage 1	intermediate	stage 2	outputs
Problem 1: turnover	all employee and job data	binary classification (supervised learning)	stable and unstable employee specs	match existing employees	stable and risky clusters
Problem 2: new hires for a certain role	employee and job data about the role, plus external candidate data	clustering*	common variable set, plus specs	match candidates	shortlist or ranking of candidates

Problem 3, 4, and 5: promotion	employee and job data about the role	binary classification (supervised learning) or parametric probabilistic modeling methods*	specs or classifier (depending on chosen method in stage 1)	match internal candidates	shortlist or ranking of candidates (depending on chosen method in stage 1)

* The issue of complementarity may or may not need (or chosen) to be explicitly addressed, depending on the choice of method.

References

Agresti, A. (1996). *An introduction to categorical data analysis* (Vol. 135). New York: Wiley.

Barney, J. B., & Hesterly, W. S. (2009). *Strategic management and competitive advantage*. Pearson Education.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information & Management*, *37*(5), 271-281.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2, No. 1). New York: Springer.

Long, L. K., & Troutt, M. D. (2003). Data mining for human resource information systems. *Data Mining: Opportunities and Challenges*, 366.

Malinowski, J., Weitzel, T., & Keim, T. (2008). Decision support for team staffing: An automated relational recommendation approach. *Decision Support Systems*, *45*(3), 429-447.